UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Cohesive Subgraph Search in Big Graphs

by Conggai Li

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Ying Zhang and Lu Qin

University of Technology Sydney
Faculty of Engineering and Information Technology

November 2020

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I, Conggai Li declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctoral of Philosophy, in the school of computer science, faculty of engineering and information technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 05/02/2021

ACKNOWLEDGEMENTS

First and foremost, I would like to deliver my sincere gratitude to my supervisor Prof. Ying Zhang for his continuous support and guidance for my PhD study and research, especially for his professionalism, patience, passion, and diligence. He is professional, efficient, patient, and diligent. His guidance broads my knowledge in computer science, improves my scientific research capacity and elevates my love for exploring the significant, undiscovered and challenging fields. Additionally, Ying is a good mentor and friend for me. Thanks to his confidence and encouragement, I am always positive and brave when experiencing challenges and even failures. This thesis could not reach its present form without his illuminating instructions.

Secondly, I would like to express my great gratitude to my co-supervisor A/Prof. Lu Qin, for his guidance and advice, especially for his support and strong confidence in me. He gave me many wonderful ideas and inspirations. His intelligence and guidance significantly extended my knowledge and helped me solve the problems I met during my study. His work efficiency gave me the hope to conduct good research without abandoning too many of other interests, which prevented me to be negative during my PhD study. Lu always has confidence in solving research problems, regardless of their complexities, which encourages me to keep

thinking and challenging myself.

Thirdly, I would like to thank Prof. Fan Zhang for his advice, especially for his guidance and help during my research study. He gave me a lot of great ideas and support during my PhD study. His guidance and help extended my knowledge and helped me solve the problems I met. I learned a lot of writing skills and many other skills for research work from Prof. Fan Zhang. His passion and brilliant ideas always inspired me during my study. Fan always inspires me for my research and gives me many great ideas for my future career. Thanks to Fan's help for this thesis.

Fourthly, I would like to thank Prof. Xuemin Lin and Prof. Wenjie Zhang for supporting the works in this thesis. I thank Prof. Lin for offering an exciting but rigorous research environment. I learned the characteristics of an excellent researcher from Prof. Lin — passion, preciseness, and earnestness. I thank Prof. Wenjie Zhang, her kindness and fantastic research works always inspired me.

Besides, I would also like to thank the following people at UNSW and UTS, Australia: Dr. Dong Wen, Dr. Ouyang Dian, Dr. Wentao Li, Dr. Xin Cao, Dr. Longbin Lai, Dr. Yixiang Fang, A/Prof. Xin Huang, for sharing your brilliant ideas and experiences. Thanks to Dr. Long Yuan, Dr. Xing Feng, Dr. Wei Li, Dr. Kai Wang, Dr. You Peng, Dr. Boge Liu, Dr. Yang Yang, Dr. Xubo Wang, Dr. Haida Zhang, Dr. Fei Bi, Dr. Chen Zhang, Mr. Xuefeng Chen, Ms. Xiaoshuang Chen, Mr. Zhengyi Yang, Mr. Qingyuan Linghu, Mr. Yuren Mao, Mr. Yixing Yang, Mr. Yu Hao, Mr. Chenji Huang, Mr. Kongzhang Hao,

Mr. Yizhang He, Dr. Bingqing Lyu, Mr. Michael Ruisi Yu, Dr. Mingjie Li, Dr. Wanqi Liu, Mr. Hanchen Wang, Mr. Yuanhang Yu, Mr. Yilun Huang, Mr. Peilun Yang, Mr. Bohua Yang, Mr. Junhua Zhang, Mr. Yuxuan Qiu, Dr. Adi Lin, Dr. Zhibin Li, Mr. Han Zheng, Mr. Huipeng Xue, Ms. Lu Zhang, Mr. Yongshun Gong, and other colleagues for sharing the happiness and bitterness with me during my PhD study. The time we spent together will never forget.

Finally, I would like to thank my parents for bringing me a wonderful life and other relatives for their understanding, encouragement, and love. Thanks to my husband, Mr. Jianjun Chen, for his company and support in my PhD study.

ABSTRACT

Graphs are widely used to model relationships in various applications, such as social science, biology, information technology, to name a few. Mining cohesive subgraphs is one of the fundamental problems in graph analytics, where the main aim is to find subgraphs with well-connected graph nodes/vertices. A variety of models have been proposed to capture the cohesiveness of subgraphs with different constraints. In this thesis, we study three cohesive subgraph models to investigate various real-life applications better.

Firstly, we would like to detect the critical users whose leave will break the user engagement of the network, i.e., lead many other users to drop out. Accordingly, we propose the collapsed k-truss problem: detect b vertices from a graph G, whose removal will lead to the smallest size k-truss, i.e., identifying some specific users to strengthen the user engagement of the network/graph. From the theoretical side, we deliver the complexity of this problem: NP-hard and inapproximate. From the practical side, we propose an efficient algorithm that can accelerate the computation by vitally reducing the number of candidates. Extensive experiments on real-life networks (graphs) demonstrate the effectiveness and efficiency of our proposed algorithm.

Secondly, we study the minimum k-core search problem. Given a graph G, an integer k and a set of query nodes $Q = \{q\}$, we aim to find the smallest size of k-core subgraph containing all the query node $q \in Q$. As one of the most representative cohesive subgraph models, k-core model has recently received sig-

nificant attention. It has been shown that this problem is NP-hard with a huge search space, and it is very challenging to find the optimal solution. There are several heuristic algorithms for this problem, but they rely on simple scoring functions, and there is no guarantee as to the size of the resulting subgraph compared with the optimal solution. Our empirical study also indicates that the size of their resulting subgraphs may be large in practice. In this thesis, we develop an effective and efficient progressive algorithm, namely *PSA*, to provide a good trade-off between the quality of the result and the search time. Novel lower and upper bound techniques for the minimum k-core search are designed. Our extensive experiments on several real-life graphs demonstrate the effectiveness and efficiency of the new techniques.

Finally, we investigate the fortress-like cohesive subgraph, p-cohesion. Morris defines the p-cohesion by a connected subgraph in which every vertex has at least a fraction p of its neighbors in the subgraph, i.e., at most a fraction $(1-p)$ of its neighbors outside. We can find that a p-cohesion ensures not only inner-cohesiveness but also outer-sparseness. The textbook on networks by Easley and Kleinberg shows that p-cohesions are fortress-like cohesive subgraphs that can hamper the cascade's entry, following the contagion model. Despite the elegant definition and promising properties, there is no existing study on p-cohesion regarding problem complexity and efficient computing algorithms to our best knowledge. In this thesis, we fill this gap by conducting a comprehensive theoretical analysis of the problem's complexity and developing efficient computing algorithms. We focus on the minimal p-cohesion because they are elementary units of p-cohesions and the combination of multiple minimal p-cohesions is a larger p-cohesion. We demonstrate that the discovered minimal p-cohesions can be utilized to solve the MinSeed problem: finding the smallest set of initial adopters (seeds) such that all the network users are eventually influenced under

the contagion model. Extensive experiments on several real-life social networks verify this model's effectiveness and the efficiency of our algorithms.

PUBLICATIONS

Fan Zhang, ***Conggai Li****, Ying Zhang, Lu Qin, and Wenjie Zhang, Finding Critical Users in Social Communities: The Collapsed Core and Truss Problems, TKDE*

Conggai Li*, Fan Zhang, Ying Zhang, Lu Qin, Wenjie Zhang, and Xuemin Lin, Efficient Progressive Minimum k-Core Search, PVLDB2020*

Conggai Li*, Fan Zhang, Ying Zhang, Lu Qin, Wenjie Zhang, and Xuemin Lin, Discovering Fortress-like Cohesive Subgraphs, revision submitted.*

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES